Cad.Est.Ling., Campinas, (43):37-54, Jul./Dez. 2002

LEARNING THE HIDDEN STRUCTURE OF SPEECH: FROM COMMUNICATIVE FUNCTIONS TO PROSODY

GÉRARD BAILLY & BLEIKE HOLM
(Institut de la Communication Parlée/INPG/Univ. Stendhal)

RESUMO Este artigo introduz um novo método, orientado via modelamento e via interação com dados comportamentais, para gerar padrões prosódicos a partir de informação metalingüística. Referimos aqui à habilidade geral da entoação de demarcar unidades de fala e veicular informação sobre as funções proposicional e interacional dessas unidades no discurso. Nossas hipóteses fortes são que (1) essas funções são diretamente implementadas como contornos prosódicos prototípicos que são co-extensivos às unidades para as quais eles se aplicam, (2) o padrão prosódico da mensagem é obtido ao superpor e adicionar todos os contornos elementares (Aubergé & Bailly, 1995). Descrevemos aqui um esquema de análise por síntese que consiste em identificar esses contornos prototípicos e separar suas contribuições respectivas nos contornos prosódicos dos dados de treinamento. O esquema é aplicado a bases de dados designadas para evidenciar várias funções entoacionais. Resultados experimentais mostram que o modelo gera contornos prosódicos adequados com pouquíssimos movimentos prototípicos.

1. INTRODUCTION

It is a commonly accepted view that prosody crucially shapes the speech signal in order to ease the decoding of linguistic and paralinguistic information by the listener. In the framework of automatic prosody generation, we aim at computing adequate prosodic parameters carrying that information. In order to automatically learn the mapping between discursive functions and prosody and eventually sketch a comprehensive model of intonation, we have to answer two main questions: *what* information is transmitted and *how* this information is encoded?

2. A MORPHOGENETIC MODEL

Encoding discourse structure – supposed to be discrete – by means of continuously varying prosodic parameters is described by a large variety of tentative approaches. A phonological interface is usually promoted that translates discourse structure in a multi-level – potentially infinite (Ladd, 1986) – phonological structure. Phonological units are typically delimited by salient prosodic events, typically accents, tones or breaks such as pauses (Hirst, Di Cristo & Espesser, 2000; Silverman *et al*., 1992). This step of phonological transfer is followed by the generation of the prosodic continuum

thanks to a specific phonetic model e.g. targets connected by interpolation functions (Hirst, Nicolas & Espesser, 1991; Pierrehumbert, 1981), series of syllable-sized contours (t'Hart, Collier & Cohen, 1990; Taylor, 2000) or superposition of contours with variable size (Aubergé, 1992; Fujisaki & Sudo, 1971; Gårding, 1991; Grønnum, 1992).
The morphogenetic model developed at ICP (Aubergé, 1992; Bailly & Aubergé, 1997) contrasts with most the models developed so far on two main points: (a) functions of discourse units are directly encoded as *global multiparametric prosodic contours* (b) the encoding of the multiple functions acting at different scopes for structuring the message is simply done by *overlapping and adding* contributions of the different contours. Our phonetic model is thus clearly global and superpositional, but contrastively with Fujisaki and Sudo (1971), the phonetic model is not motivated by a production mechanism – although this mechanism may have acted as a bootstrap – but by communication needs, i.e. maintaining perceptual contrasts that ensure optimal decoding of the functions.
Note that we have added in the current implementation of the model another strong hypothesis to the point (a): the global contours are only parameterized by the *scope* – or domain – of the function, i.e. the size of the units the function is applied to, and does not depend on the nature and internal organization of the units.

2.1. Multiparametric characterization of prosody

We must also point out that we generate *multiparametric prosodic contours* i.e. melody *and* rhythmic organization of the synthetic message are generated together within the same generation process as amplified in Figure 1. In fact each *Inter Perceptual-Center Group* (IPCG) (Barbosa & Bailly, 1994a) is characterized by a melodic contour (stylized by three F0 values on the vocalic nucleus) and a lengthening factor (that will stretch or compress the segmental constituents in a nonlinear way). This has been made possible by the work of P. Barbosa (Barbosa & Bailly, 1994a; Barbosa & Bailly, 1994b; Barbosa & Bailly, 1997) on macrorhythm, giving access to a speech tempo parallel to the melodic curve. Morlec (Lorlec, Aubergé & Bailly, 1995; Morlec, Bailly & Aubergé, 1995; Morlec, Bailly & Aubergé, 1996; Morlec, Bailly & Aubergé, 1997) first implemented this multipametric generation scheme.
An extensive study of the perceptual impact of F0 stylization has been conducted by S. de Tournemire (1994). We choose to characterize the melodic curve by 3 F0 values per GIPC respectively at 10%, 50% and 90% of the vocalic nucleus. This simple strategy explains the oscillations exhibited by the prosodic contours due to the adjacent consonantal dips that a smoothing procedure (such as proposed in Grønnum, 1992) could easily wipe out. Note that we compensate at synthesis time this crude stylization by adding to the final melodic contour the residual F0 trajectories of the concatenated segments (here polysounds, Bailly, Barbe & Wang, 1992) obtained by the same stylization procedure applied to the carrying words (here logatoms) from which the segments are extracted. The stylization procedure gives the melodic skeleton and the segments give the flesh that is glued on the skeleton. Note that this generation

process is entirely compatible with a superposition model. If the melodic skeleton is produced by a global approach involving the superposition of dynamic prosodic prototypes (see below) and the flesh is given by a lexicon lookup, both are overlapped and added to produce the final contour.

The segmental durations are obtained by a multi-level timing generation process similar to Campbell (1992) but using the IPCG as an intermediate rhythmical unit. Each IPCG is characterized by a lengthening/shortening factor equal to the quotient between the actual duration of the ICPG and an expected ICPG duration. This expected duration is a weighted sum of (a) the sum of the mean values of its constitutive segments (b) the average duration of an ICPG comprising n segments.

A z-scoring procedure is then applied in order to distribute the actual ICPG duration among its constitutive segments. Pause insertion is obtained by saturating the lengthening factor of the IPCG: the pause duration is computed as the duration loss between the desired lengthening factor and the saturated lengthening factor (for further details please refer to Barbosa & Bailly, 1997). Thus contrary to prosodic phonology, pause is an emergent process resulting from low-level constraints (overall speech rate, pausing strategy resulting from the control of the saturation curve) and do not determine a priori the performance structure.

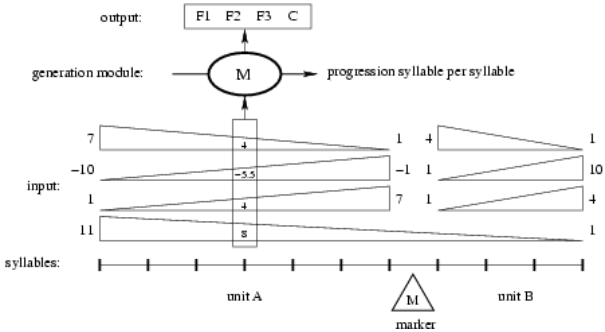

Figure 1: M is a contour generator that converts linear ramps – anchored on the boundaries of units A and B – into prosodic trajectories: for each syllable of the units, it delivers three F0 values (F0 values at 10%, 50% and 90% of the vocalic nucleus of each syllable) and a lengthening factor (phoneme durations are further computed together with pause generation using the procedure described in Barbosa & Bailly, 1997).

2.2. Contour generators

Each discourse function may be applied to diverse discourse units. We define the *scope* of a function as the continuous set of words which are concerned with this function. These functions typically assign a communicative value to a unit or qualify the link between units within the discourse. The *segmentation* function can for example indifferently demark a word, a group or a clause off the utterance. The same *qualification* function is applied indifferently to an adjective, a noun complement or a clause qualifying a preceding noun or nominal group (cf. 3.3.2). Similarly an *emphasis* function could be indifferently applied to any constituent of the discourse.

Each discourse function is then encoded by a specific prototypical contour anchored to the function's scope by so-called *landmarks*, i.e. beginning and end of the units concerned with this function. As the discourse function can be applied to different scopes, it is characterized by a family of contours – some sort of prosodic "clichés" (Fónagy, Bérard & Fónagy, 1984).
General-purpose *contour generators* have been developed in order to be able to generate a coherent family of contours given only their scope. These contour generators are actually implemented as simple feedforward neural networks (Holm & Bailly, 2000) receiving as input linear ramps giving the absolute and relative distance of the current syllable from the closest landmarks and delivering as output the prosodic characteristics for the current syllable (see Figure 2). Each network have very few parameters – typically 4 input, 15 hidden and 4 output units = 4*(15+1)+15*(4+1) = 139 parameters – to be compared to the thousands parameters necessary to learn a "blind" mapping between phonological inputs and prosodic parameters such as in (Chen, Hwang & Wang, 1998; Traber, 1992). We have shown that our contour generators implement a so-called Prosodic Movement Expansion Model (PMEM) that describes how prototypical contours develop according to the scope (see for example Figure 2): the set of prototypical contours that a contour generator implementing a certain function actually generates is called in the following a *dynamical prototype*. Note that the choice of the neural networks implementation of the PMEM is not exclusive, but offers an efficient learning paradigm as described below.
The final multiparametric prosody is thus obtained by superposing and adding the many contours produced by a few independent contour generators (typically 3 or 4) and parameterized by their smaller or larger scopes.

2.3. Analyzing prosody

The mapping between discourse structure and the phonological structure is usually not straightforward: a direct mapping between these two structures is highly problematic (Marsi *et al.*, 1997; Pierrehumbert & Hirschberg, 1990). In fact, while phonological structures are often represented as tree structures (Ladd, 1986; Ladd, 1988; Selkirk, 1984), phonological events quite distant in time may act as a whole – although belonging to distinct phrases - and should be linked by additional semantic links (Marsi *et al.*, 1997) that makes the phonological structure complex and often violate the hypothesis that rules the geometry of phonological trees Most authors thus rely on a specific analysis technique – often requiring expertise – for constructing first a phonological tree from raw acoustic data. Then a further mapping between this surface phonological tree and communicative functions is required. As amplified in the introduction of this chapter, we voluntarily skip this step of converting the discourse structure into a deep or a surface phonological representation: our superpositional phonetic model implements directly the diverse communicative functions the message is supposed to carry via the superposition of multiparametric prototypes.
The problem is now to recover these multiparametric prototypes from raw data. In the case of a superpositional model, the problem is often ill-posed since each observation

is in general the sum of several contributions, i.e. here the outputs of contributing contour generators. We thus need extra constraints to regularize the inversion problem, e.g. shapes/equations of the superposed components as in (Mixdorff, 2000). In our phonetic model, shapes of the contributing contours are *a priori* unconstrained – which we feel to be important in a first time since we have shown that contours may potentially have complex shapes (e.g. those encoding attitudes at the sentence level in Morlec, Bailly & Aubergé, 2001). Note however that nothing forbids in the following framework to later add constraints (such as imposing exponential shapes as in the Fujisaki's model) on those contours that are well understood in order to ease the emergence of other contours.

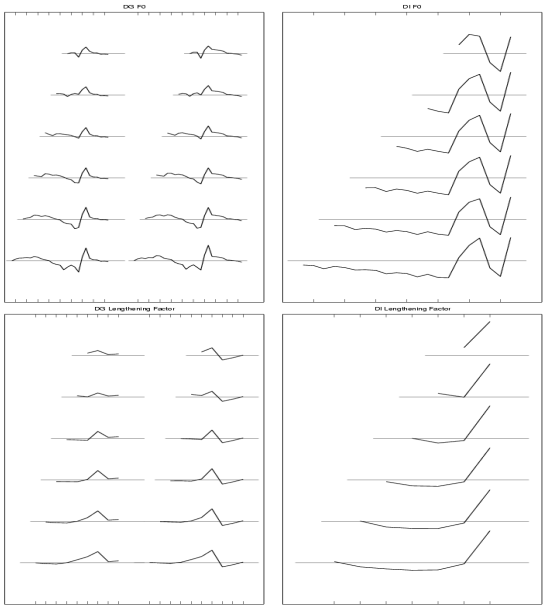

Figure 2: Expansion of the prosodic contour produced by a contour generator encoding different functions: (a) a presupposition relationship between two units. (b) an incredulous question on a sentence of 2 to 7 syllables. Top: melodic prototypes; bottom: lengthening factor profiles. In (a) the length of the first unit is varied from 2 to 7 syllables while the second unit has 2 (left column) and 3 (right column) syllables.

The shapes of the contributing contours emerge here as a by-product of an inversion procedure that parameterize contour generators in such a way that the prosodic contours predicted by overlapping and adding their contributions in the discourse best predicts observed realizations. The analysis procedure is by essence reversible and our phonological model – implemented as dynamical prototypes – emerges from an iterative analysis-by-synthesis process as follows:

1. we generate the assumed contribution of each discourse function at each supposed scope in the corpus with the generators. In their initial state, they produced a null output.
2. we compute a prediction error by subtracting the sum of these elementary overlapping contours to the original prosodic contours observed in the corpus.
3. this prediction error is then distributed on the contributing contours, i.e. partial contributions are added to them assuring that the superposition of these adjusted contours equals the observed contour for each sentence. For now, we use a simple repartition scheme consisting in dividing equally – for each syllable - the prediction error between contributing contours.
4. these new contours are used as targets during a classical learning procedure for neural networks.

These steps are iterated until the prediction error of a test set reaches a minimum. This scheme relies on three hypothesis:

a. the prediction error contains the information that is contained in natural prosody but not (yet) captured by the contributing generation modules;
b. step 4 provides a filter capturing regularities within each target set, i.e. if a contribution of the prediction error is attributed to the "wrong" module, it should have no systematic relation to the associated input values and will thus be flattened;
c. the family of contours that contour generators are able to produce is finite i.e. the simple phonotactic information provided to the contour generators constrains the topology of the mapping of the generators. Our implementation as neural networks seems suited since it fulfills the conditions (b) and (c), but – as stated above – other choices are possible.

Table 1: *RMS prediction errors (correlation coefficients) for different corpora. F0 errors are given in semitones, IPCG and phoneme durations in ms. The last column gives the number of syllables and phonemes considered. The last syllables of the sentences are excluded.*

	F0 [st]	IPCG [ms]	Phon. [ms]	Nsyl / Nphon
Math	2.29 (0.87)	105 (0.90)	31.2 (0.67)	2805 / 7557
DC	1.89 (0.81)	34.6 (0.86)	22.8 (0.74)	1726 / 3868
DI	1.44 (0.94)	27.6 (0.87)	17.2 (0.73)	849 / 1964
QS	1.31 (0.67)	28.5 (0.87)	17.6 (0.71)	1120 / 2581
EV	2.09 (0.91)	28.1 (0.87)	16.7 (0.74)	1005 / 2309

EX	2.88 (0.80)	28.1 (0.89)	18.5 (0.72)	1005 / 2322
SC	1.17 (0.65)	27.5 (0.87)	16.9 (0.73)	1000 / 2297
Text	1.37 (0.77)	21.5 (0.86)	15.1 (0.79)	11210/23239

3. APPLYING THE MORPHOGENETIC MODEL TO DIVERSE CORPORA

We summarize here the results obtained on different corpora using half of the corpus as learning data. The prediction statistics are given in Table 1 using all available data.

3.1. Maths

3.1.1. The corpus

The Math corpus (Holm, Bailly & Laborde, 1999) was established in order to study how prosody may encode highly embedded dependency relations between constituents of an utterance. Read Mathematical Formulae (MF) were chosen because they offer a deep syntactical structure and because they are – when spoken – often ambiguous, forcing thus the speaker and the listener to use prosodic cues. All formulae are algebraic equations such as proposed in 4th grade exercises. They involve classical operations on 2nd degree polynomials. The corpus was generated automatically by systematically varying the length and syntactic depth of constituents. We end up with 157 MF that were recorded by one male French speaker who was instructed not to use lexical structural markers – as "open parenthesis" – but to make use of prosody.
Each formula has been uttered twice. In order to describe the natural variability of our data we give here RMS-errors (correlations) between the repetitions: phoneme durations: 0.857/20.6 ms, IPCG durations: 92.6ms (0.919) and F0: 2.0 semi-tones (0.902). The two versions have 579 – internal – pauses in common of a total of 616. Pause durations[1] are correlated by 0.917. Note that even in case of such a close repetition, we still have a large variance. These values serve as reference for the corresponding values between the model's predictions and the original variance given Table 1.

[1] A pause which is not realized in either stimuli is considered with null duration.

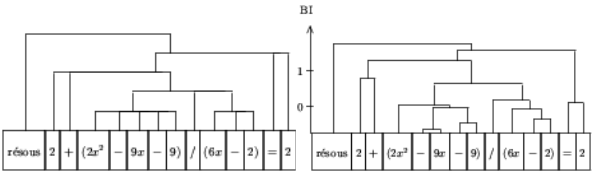

Figure 3: Comparing the syntactic structure of a MF (left) with the performance structure of its spoken form (right). Note the decrease of the boundary strength between equals (=) and its right operand (2), cueing the tendency to group small accentual units.

3.1.2. Discursive functions

Performance structures (a tree-representation of word-final lengthening factors Gee & Grosjean, 1983; Grosjean, Grosjean & Lane, 1979; Monnin & Grosjean, 1993) of spoken maths (Holm, Bailly & Laborde, 1999) reflect nicely the underlying syntactic structure of the MF with the tendency to balance the strengths of boundaries across operators according to the relative syllabic weights of left and right operands (see Figure 3). For example, operators tend to group with the smallest operand, tendency already mentioned by Campbell (1993) for junction words like prepositions.
We thus decided to use here only three basic communication functions:

1. Introduced imperative statement: all sentences have the form (*solve*)M(*MF*)
2. Linking left operand with operator e.g. (*9x*)L(*/(6x+3)*)
3. Linking operator with its right operand e.g. (*/*)R((*6x+3)*)

Each complex formula is thus decomposed into sets of embedded dependency relations. The MF of Figure 4 is thus decomposed into 1 M, 7 R and 6 L relations between various units as below:

(*Résouds*)M(*valeurabsoluede5x+2sur8x-6+9xsur6x+6>2*)
(*valeurabsoluede5x+2sur8x-6+9xsur6x+6*)L(*>2*) (>)R(*2*)
(*valeurabsoluede*)R(*5x+2sur8x-6+9xsur6x+6*)
(*5x+2sur8x-6*)L(*+9xsur6x+6*) (+)R(*9xsur6x+6*)
(*5x+2*)L(*sur8x-6*) (*sur*)R(*8x-6*)
(*9x*)L(*sur6x+6*) (*sur*)R(*6x+6*)
(*8x*)L(*-6*) (-)R(*6*)
(*6x*)L(*+6*) (+)R(*6*)

Note that this description contains no explicit notion of the hierarchical level of a discourse function. Nevertheless, a hierarchical structure may emerge since high level functions have bigger scopes.

3.1.3. Predictions

The z-score repartition scheme (Barbosa & Bailly, 1997) generates 558 pauses at a location common to either natural versions, omit 58 pauses and generates 128 extra pauses[2]. Durations are correlated by 0.794. The results for F0 and syllable durations are close to the natural variability. The rather big difference in phoneme durations (105ms) is considered to be perceptually less crucial (the model predicts in fact only Inter-Perceptual Centers (IPCG) durations: phoneme and pause durations are obtained by a z-scoring procedure (Barbosa & Bailly, 1994b; Barbosa & Bailly, 1997) not optimized for that particular speaker: the repartition algorithm actually tends to privilege the generation of silent pauses versus stretching segments.

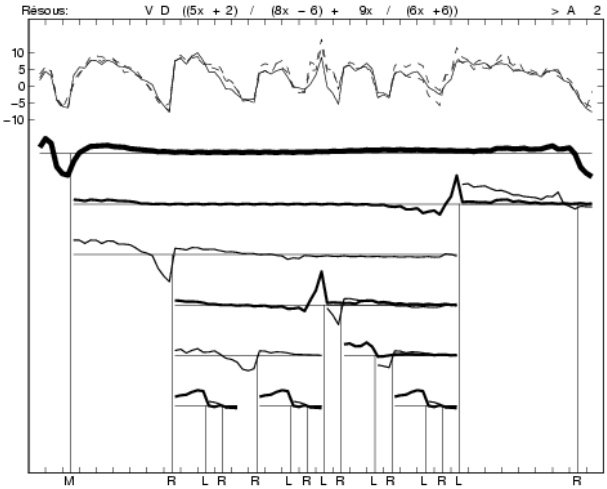

Figure 4: Predicting/analyzing the melody of a complex formula as the superposed contributions of three contour generators encoding an introduced assertion (M) and two dependency relations between the left operand and the operator (L) and between the operator and its right operand (R). Top are superposed the prediction (plain) and the original F0 stylization (dashed). M contribution (thick) is shown below with L (plain) and R (thin).

3.2. Prosodic attitudes

3.2.1. The corpus

The corpus of prosodic attitudes (Morlec, Bailly & Aubergé, 2001) reveals the existence of statistically significant global prosodic contours that encode communicative functions at the utterance-level: 322 syntactically balanced unmarked sentences were uttered by one speaker with six different prosodic attitudes: declarative (DC), question (QS), exclamation (EX), incredulous question (DI), suspicious irony

[2] These locations and associated pause durations tend globally to enhance the phrasing structure of the utterance such as disconnecting the right operand from the major operand “equals to”.

(SC) and obviousness (EV). The morpho-syntactic structure of the sentences and their lengths (between 1 and 8 syllables) were systematically varied in order to eliminate coincidental covariations between the contours encoding the communicative function at the utterance-level and the morpho-syntactic structure of the sentence.

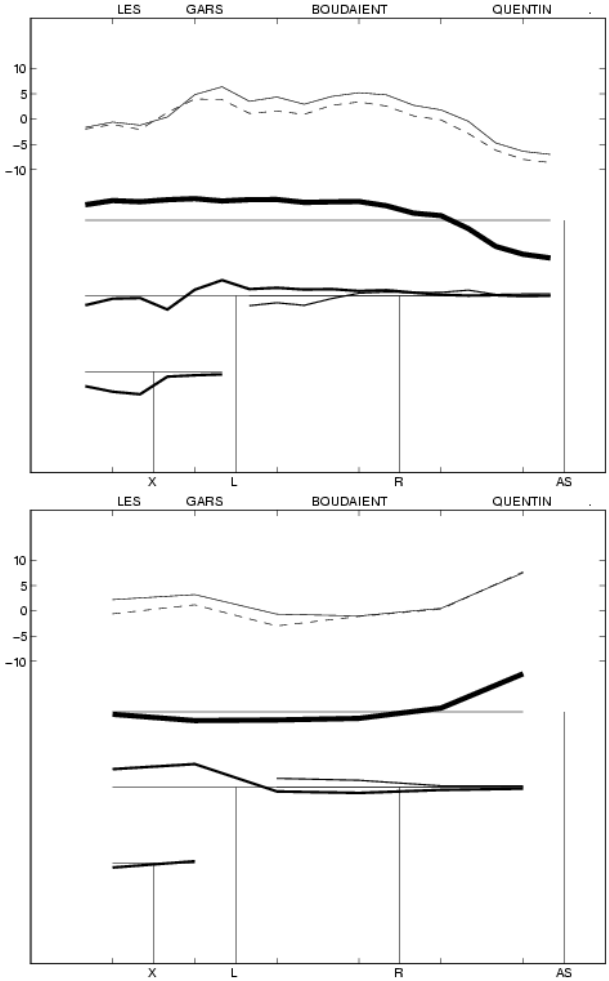

Figure 5: Predicting/analyzing the melody (top) and syllable lengthening (bottom) of a statement.

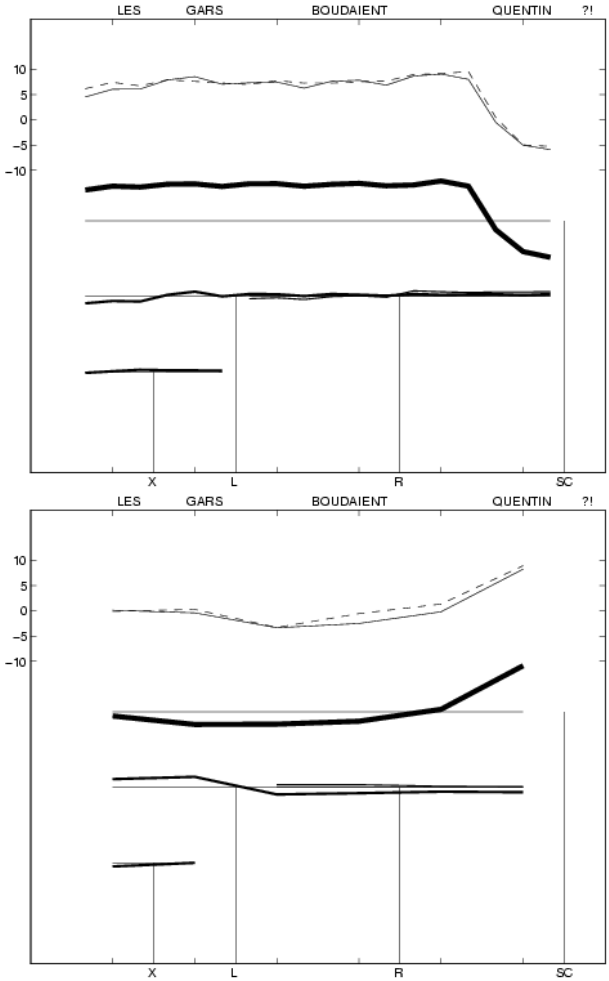

Figure 6: Predicting/analyzing the melody (top) and syllable lengthening (bottom) of a sentence uttered with a suspicious irony. Note that the amplitudes of the contours carrying phrasing structure are quite reduced compared with Figure 5.

3.2.2. Discursive functions

Besides the obvious encoding of the prosodic attitude whose scope is clearly the whole utterance with no internal landmark, we added discursive markers for encoding the morpho-syntactic structure of the sentences. As in maths, operators here are the governors of each group i.e. the verb in a verbal group, the noun in a nominal group, etc…We distinguished four discursive functions:

1. Prosodic attitude operating at the sentence level. Tags are associated with each attitude (DC, QS, EX, DI, SC, EV)
2. Linking units left to the governor: a L tag is introduced between the left unit and its governor.
3. Linking units right to the governor. In French most qualifications are right to the governor: only a few adjectives are positioned before the noun. As amplified in the §0, we do not distinguish between the different units that could address the same function e.g. qualifying the noun with a simple adjective, an adjective group, a noun complement or a qualificative clause): a R tag is introduced between them.
4. Linking function words to the proper unit (see discussion below): a X tag is introduced between them.

Each sentence is thus decomposed into sets of embedded dependency relations. The sentence of Figure 6 is thus decomposed into 1 SC, 1 R, 1L and 2X relations between various units as below:

SC(*Les gamins coupaient des rondins*)
(*Les gamins*)L(*coupaient des rondins*)
(*coupaient*)R(*des rondins*)
(*Les*)X(*gamins*)
(*des*)X(*rondins*)

3.2.3. Predictions

The decomposition of utterances into sentential and phrasal intonation is performed for each prosodic attitude separately. A further analysis demonstrates that contours carrying morpho-syntactic information are quite reduced especially for non modal attitudes (see Figure 6). In this case, the speaker is supposed to doubt, be ironical or suspicious about a previous assertion of his interlocutor, who does not require phrasing to be returned back to him. The overall flat pattern of these contours explains the rather low correlation of F0 for QS and SC. The very small RMS-errors for these attitudes indicate that the predictions are nevertheless suitable. The biggest errors of FO are found for EV and EX – they are mainly due to emphatic accents not (yet) modeled.

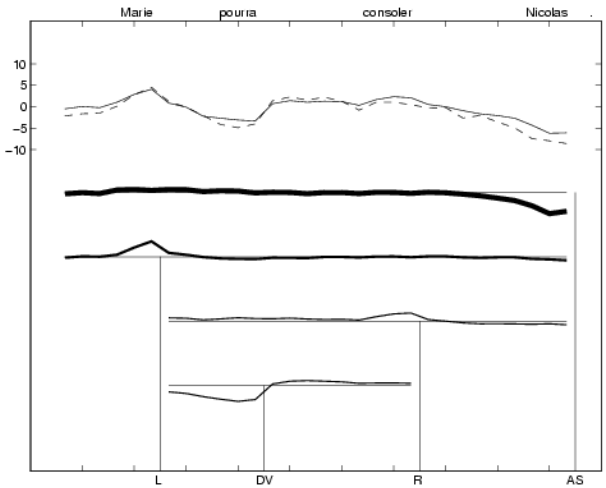

Figure 7: Predicting/analyzing the melody of a read sentence "Marie pourra consoler Nicolas". The DV contour is often observed and is generally used to segment determinants or auxiliaries, here "pourra" from "consoler"

3.3. Text reading

This eclectic studies of highly dedicated materials assess some properties of the natural intonation and evidence some important features of the morphogenetic model:

a. the existence of global contours that encapsulate co-occurring salient event. These analysis results together with gating experiments (Aubergé, Grépillat & Rilliard, 1997) confirm the pertinence of our Gestalt approach.
b. the possibility of intonation – with syntax – of carrying structural information with very few contour generators.

This should however not obscure the main technological grail of speech synthesis: being able to read texts.

3.3.1. The corpus

A corpus of 1000 sentences (between 4 and 20 syllables) was recorded by a female French speaker. This corpus was designed to cover extensively the standard declarative form of French sentences NP VP, while extending NP from a simple pronoun to a complex nominal group with adjectives, noun complements and simple qualificative clauses, and VP with adverbs or verb complements. We were particularly interested in assessing this independence between the nature and internal organization of the units and the functions they entertain with each other.
We used the same assumptions as previously for decomposing each sentence into embedded units. The systematic opposition in the corpus between a full verb and a

modal auxiliary+infinitive reveals the necessity to introduce an additional function DV to segment between the modal auxiliary and the infinitive. The similarity between PMEMs produced by X and DV leads us to further assimilate X and DV functions as a general function used to segment between a function word and the content word it introduces.

3.3.2. Predictions

This corpus yields the smallest prediction errors – as well for F0 as for durations (apart from the already discussed QS and SC special cases).
Figure 7 illustrates the hypothesized independence between internal organization of units and their functional role. The four sentences were chosen to employ the same functions with more or less identical scopes. The contours show that the nature of the qualifying part in the Gn (adjective: “menaçant”; noun complement: “de son mas”, “de mes maisons” or qualicative clause: “qui veut manger”) does not change significantly the overall shape of the contours. The persisting differences may be understood as modulations interior to the qualifying unit – e.g. due to contours of type X (cf. 3.2.2).

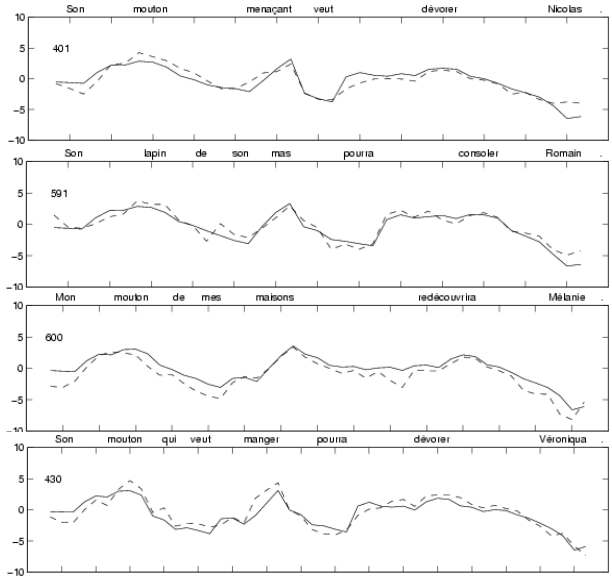

Figure 7: Varying the internal structure of a unit with a given functional relation doesn’t change the overall contour-shape.

4. CONCLUSIONS AND PERSPECTIVES

The analysis-by-synthesis procedure presented here gives access to the *hidden structure* of intonation: the phonetic implementation of discourse functions emerges

from the automatic parameterization of contour generators. This procedure is data-driven but also model-constrained and thus converges towards optimal prototypical contours that satisfy *both* bottom-up (close-copy synthesis) and top-down (coherent phonological description) constraints.

Such a phonology of prototypes can easily include a paradigm for learning automatically *alloprosodic* variations i.e. privileged directions of variations around the prototypes and implement a model of phonological *gradience* (Gussenhoven, 1999) able to encode and modulate the degree of importance of the information carried by the contour in the discourse.

By applying the model to different communicative functions we have demonstrated that this model can actually capture statistically significant prosodic variations with a rather few number of prototypical movements and that it generates faithful and varied prosodic contours. This model provides a useful tool for analyzing the "hidden" structure of intonation i.e. decomposing a surface prosodic contour into overlapping contours that actually implement a given communicative function in a statistically-significant way. We plan to exploit this model for analyzing multilingual corpora and implementing new functions. For instance, we are currently working on Galician, a language with lexical stress.

We have also tried to demonstrate that this model-based comprehensive generation scheme may be compatible with a certain technological efficiency: confronting data-driven models against such thematic databases used here should provide an interesting basis of comparison between models and approaches that we are still looking for.

ACKNOWLEDGEMENTS: This work was supported by Cost258. We thank H. Loevenbruck and G. Rolland for providing us the text reading corpus and fruitful comments on the first version of this paper. A special thank also to our proof reader L. Ménard.

REFERENCES

AUBERGÉ, V. (1992). Developing a structured lexicon for synthesis of prosody. In BAILLY, G. & BENOÎT, C. (eds.). *Talking Machines: Theories, Models and Designs*. Elsevier B.V. pp. 307-321.

AUBERGÉ, V. & BAILLY, G. (1995). Generation of intonation: a global approach". In Proceedings of the European Conference on Speech Communication and Technology. Madrid. pp. 2065-2068.

AUBERGÉ, V.; GRÉPILLAT, T. & RILLIARD, A. (1997). Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. In Proceedings of the European Conference on Speech Communication and Technology. Rhodes - Greece. pp. 871-874,.

BAILLY, G. & AUBERGÉ, V. (1997). Phonetic and phonological representations for intonation. In Progress in Speech Synthesis. VAN SANTEN, J.P.H.; SPROAT, R.W.; OLIVE, J.P.; HIRSCHBERG, J. (Eds.). New York: Springer-Verlag. pp. 435-441.

BAILLY, G.; BARBE, T. & WANG, H. (1992). Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system. In Talking Machines: Theories, Models and Designs. G. BAILLY; C. BENOÎT (eds.). Elsevier B.V. p. 323-333.

BARBOSA, P. & BAILLY, G. (1994a). Characterisation of rhythmic patterns for text-to-speech synthesis. Speech Communication. 15: 127-137.
_____. (1994b). Generating pauses within the z-score model. In Proceedings of the ETRW on Speech Synthesis. New Paltz , USA. pp. 101-104.
_____. (1997). Generation of pauses within the z-score model. In VAN SANTEN, J.P.H.; SPROAT, R.W.; OLIVE, J.P.; HIRSCHBERG, J. (eds.). Progress in Speech Synthesis. New York: Springer-Verlag. pp. 365-381.

CAMPBELL, W. N. (1992). Multi-level timing in speech. Unpublished PhD thesis. Brighton, UK: University of Sussex.
_____. (1993). Automatic detection of prosodic boundaries in speech. Speech Communication. 13: 343-354.

CHEN, S.-H.; HWANG, S.-H. & WANG, Y.-R. (1998). An RNN-based prosodic information synthesizer for Mandarin text-to-speech. Speech and Audio Processing. 6 (3): 226-239.

DE TOURNEMIRE, S. (1994). Recherche d'une stylisation extrème des contours de F0 en vue de leur apprentissage automatique. In Journées d'Etudes sur la Parole. Trégastel, France. pp. 75-80.

FÓNAGY, I.; BÉRARD, E. & FÓNAGY, J. (1984). Clichés mélodiques. Folia Linguistica. 17: 153-185.

FUJISAKI, H. & SUDO, H. (1971). A generative model for the prosody of connected speech in Japanese. Annual Report of Engineering Research Institute. 30: 75-80.

GÅRDING, E. (1991). Intonation parameters in production and perception. In Proceedings of the International Congress of Phonetic Sciences. Aix-en-Provence, France. pp. 300-304.

GEE, J.-P. & GROSJEAN, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. Cognitive Psychology. 15: 411-458.

GRØNNUM, N. (1992). The ground-works of Danish intonation.Copenhagen: Museum Tusculanum Press - Univ. Copenhagen.

GROSJEAN; GROSJEAN, F. & LANE (1979). The Patterns of Silence: Performance Structures in Sentence Production. Cognitive Psychology. 11: 58-81.

GUSSENHOVEN, C. (1999). Discreteness and gradience in intonational contrasts. Language and Speech. 42: 283-305.

HIRST, D.; NICOLAS, P. & ESPESSER, R. (1991). Coding the F0 of a continuous text in French: an experimental approach. In Proceedings of the International Congress of Phonetic Sciences. Aix-en-Provence, France. pp. 234-237.

HIRST, D.J.; DI CRISTO, A. & ESPESSER, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In Prosody: Theory and Experiment, M. HORNE (ed.). Dordrecht, The Netherlands: Kluwer Academic Publishers. pp. 51-87.

HOLM, B. & BAILLY, G. (2000). Generating prosody by superposing multi-parametric overlapping contours. In Proceedings of the International Conference on Speech and Language Processing. Beijing, China. pp. 203-206.

HOLM, B.; BAILLY, G. & LABORDE, C. (1999). Performance structures of mathematical formulae. In *Proceedings of the International Congress of Phonetic Sciences*. San Francisco, USA. pp. 1297-1300.

LADD, D.R. (1986). Intonational phrasing: the case for recursive prosodic structure. *Phonology Yearbook*. 3: 311-340.

_____. (1988). Declination "reset" and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*. 84 (2): 530-544.

MARSI, E.C.; COPPEN, P.-A. J.M.; GUSSENHOVEN, C.H.M. & RIETVELD, T.C.M. (1997). Prosodic and intonational domains in speech synthesis. In VAN SANTEN, J.P.H.; SPROAT, R.W.; OLIVE, J.P.; HIRSCHBERG, J. (eds.). *Progress in Speech Synthesis*. New York: Springer-Verlag. pp. 477-493.

MIXDORFF, H. (2000). A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters. In *International Conference on Acoustics*, Speech and Signal Processing. Istanbul, Turkey. pp. 1281-1284.

MONNIN, P. & GROSJEAN, F. (1993). Les structures de performance en français: caractérisation et prédiction. *L'Année Psychologique*. 93: 9-30.

MORLEC, Y.; AUBERGÉ, V. & BAILLY, G. (1995). Evaluation of automatic generation of prosody with a superposition model. In *Proceedings of the International Congress of Phonetic Sciences*. Stockholm, Sweden. pp. 224-227.

MORLEC, Y.; BAILLY, G. & AUBERGÉ, V. (1995). Synthesis and evaluation of intonation with a superposition model. In *Proceedings of the European Conference on Speech Communication and Technology*. Madrid, Spain. pp. 2043-2046.

_____. (1996). Generating intonation by superposing gestures. In *Proceedings of the International Conference on Speech and Language Processing*. Philadelphia, USA. pp. 283-286.

_____. (1997). Synthesising attitudes with global rhythmic and intonation contours. In *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes, Greece. pp. 219-222.

_____. (2001). Generating prosodic attitudes in French: data, model and evaluation. *Speech Communication*. 33 (4): 357-371.

PIERREHUMBERT, J. (1981). Synthetizing intonation. *Journal of the Acoustical Society of America*. 70 (4): 985-995.

PIERREHUMBERT, J. & HIRSCHBERG, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P.R. COHEN; J. MORGAN & M.E. POLLAK (eds.). *Intentions in Communication*. Cambridge, MA: MIT Press. pp. 271-311.

SELKIRK, E.O. (1984). *Phonology and Syntax*. Cambridge, MA: MIT Press.

SILVERMAN, K.; BECKMAN, M.; PITRELLI, J.; OSTENDORF, M.; WIGHTMAN, C.; PRICE, P.; PIERREHUMBERT, J. & HIRSCHBERG, J. (1992). TOBI: a standard for labeling English prosody. *International Conference on Speech and Language Processing*, v. 2, 867-870.

T' HART, J.; COLLIER, R. & COHEN, A. (1990). *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

TAYLOR, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*. 107 (3): 1697-1714.

TRABER, C. (1992). F0 generation with a database of natural F0 patterns and with a neural network. In G. BAILLY; C. BENOÎT (eds.). *Talking Machines: Theories, Models and Designs*. Elsevier B.V. pp. 287-304.